Using rules for assessing and improving data quality: A case study for the Norwegian State of Estate report

Ling Shi[1] and Dumitru Roman[2]

[1]Statsbygg, Pb. 8106 Dep, 0032 Oslo, Norway
ling.shi@statsbygg.no
[2]SINTEF, Pb. 124 Blindern, 0314 Oslo, Norway
dumitru.roman@sintef.no

Abstract. The Norwegian State of Estate (SoE) report service – a service providing information about central government properties in Norway – is a result of integrating cross-domain government data originating from the Norwegian cadastral system, Business Entity Register, Building Accessibility Register and Statsbygg's property management system. This paper presents a rule-based approach to assess and improve the quality of the data upon which the SoE service is built. The approach develops a set of rules to specify a common data schema, rules for data quality assessment, and three dedicated measurement metrics for data integration. Application scenarios of the approach in identifying data inconsistencies in the sources are exemplified with strategies to improve data quality.

Keywords: Rule-based approach, Data quality assessment, Data integration, Report service

1 Introduction

A State of Estate (SoE) report[1] produces a complete list of state-owned real estates[2], and represents a key input to the decision making process of the government or other stakeholders to increase the effectiveness of the public resources allocation. The SoE report in Norway is published as an attachment[3] to the proposed parliamentary resolution No.1 every four years by Statsbygg[4] on behalf of the Ministry of Local Government and Modernization[5]. The current reporting process is manual, static and error-prone and the report is outdated when it is produced, therefore the report does not

[1] An example of such a State of Estate report from the UK government can be found at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/200448/SOFTE2012_final.pdf.

[2] Real estates can also be called real properties, properties or cadastral parcels if the properties are registered at the national cadastral system.

[3] https://www.regjeringen.no/contentassets/f4346335264c4f8495bc559482428908/no/sved/stateigedom.pdf

[4] http://www.statsbygg.no/Om-Statsbygg/About-Statsbygg/

[5] https://www.regjeringen.no/en/dep/kmd/id504/

properly support the decision making process. A new State of Estate (SoE) report generation process was introduced in [1] to carry out the reporting task in a more effective way, realized as a reporting service. It aims to provide users with a dynamic and up-to-date report, including data visualization features, to better support the users' decision making process.

The new SoE reporting service reuses existing government data, from both open and proprietary sources, and integrates them in a way that can serve as a basis for the creation of the SoE service. The data sources include the Norwegian cadastral system[6], Business Entity Register[7], Building Accessibility Register[8], and Statsbygg's property management system. Though data are collected from the most authoritative government agencies, they are not 100% consistent with each other and the inconsistency is one of the main challenges to create the SoE service. Our focus and contribution in this paper is to establish a rule-based approach which develops a set of rules to assess and improve the data quality. A rule-based approach is suitable in this context, quick to implement, and easy to document and understand.

The rest of this paper is structured as follows. Section 2 describes the SoE report service case focusing on the value proposition. Section 3 presents the rule-based approach for data quality assessment and improvement. Section 4 summarizes the paper and outlines possibilities for further work.

2 Norwegian SoE value proposition

The State of Estate (SoE) service is a reporting service for state-owned properties in Norway. The customers of the service include:

- Ministry of Local Government and Modernization (KMD);
- Property owners in the public sector;
- Public audience including the media;
- Real estate development companies.

The SoE service allows the property owners in the public sector to do quality assessment [2] on data of their own real estates. It should also provide the reporting and visualization functions of state-owned properties to the above mentioned customer groups.

The value proposition canvas[9] for property owners in the public sector is shown as an example in Fig. 1 and Fig. 2. The property owners' pains and gains are listed up in the customer segment profile. Improved data quality, reliability, completeness and accessibility are the main gains against the pains on static reports, manual data collection and quality control and missing records. The value proposition map designs the SoE report service and its Gain Creators and Pain Relievers, including data quality

6 http://www.kartverket.no/en/Land-Registry-and-Cadestre/

7 https://www.brreg.no/home/

8 https://byggforalle.no/uu/sok.html?&locale=en

9 https://strategyzer.com/canvas/value-proposition-canvas

requirements on improved quality and completeness of the report and reduced number of missing buildings.

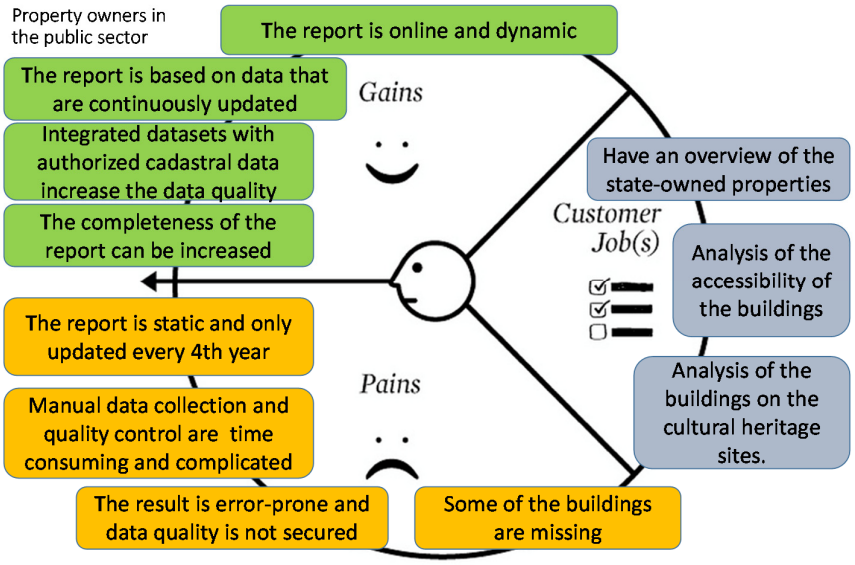

Fig. 1. Value proposition canvas – customer segment profile

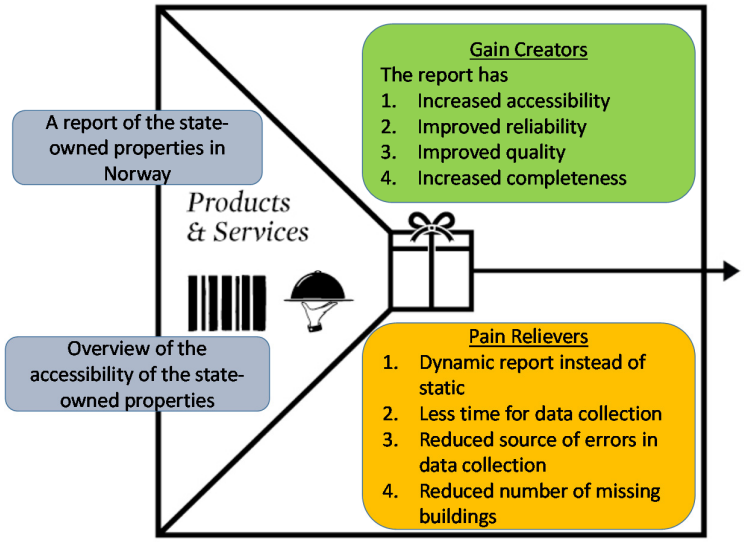

Fig. 2. Value proposition canvas – value proposition map

3 Rule-based data quality assessment and improvement

In order to meet the data quality requirements illustrated in the value proposition canvas, we established a rule-based approach to assess and improve the data quality firstly of the source data and thereafter the result data of integration. Data quality rules are contextual [3] and the rules developed in this section are therefore valid within the context of the SoE report service though the general method to categorize and define the rules can be reusable in other contexts. The following sub-sections cover the rules to specify a common data schema in Sub-section 3.1, rules for data quality assessment in Sub-section 3.2, measurement metrics for data integration is introduced in Sub-section 3.3, and strategies for improving data quality is presented in Sub-section 3.4.

3.1 Rules to specify a common data schema

Data inconsistency and redundancy are well-known challenges in cross-domain data integration. For example, the cadastral ownership information and building information are registered both in the cadastral system and Statsbygg's property management system with different updating status; a property owner's organization number and name are registered both in the cadastral system and Business Entity Register but the cadastral system and the Business Entity Register are not synchronized. This subsection presents several steps: firstly to decide the master source systems for the involved domains, afterwards to define rules to specify a common data schema and integration keys.

As a first step we make a decision on which source system is the master for each domain or sub-domain involved in the data integration. The government organizational structure reflects the domain responsibility for government data. Both the Business Entity Register and the Cadastral system are national registers and provide data with relatively high quality, therefore those two systems are defined as the master or primary data sources for the correspondingly organization domain and cadastral domain. Statsbygg's property management system is defined as a supplementary data source for the cadastral domain. The Building Accessibility Register is defined as a supplementary data source for the cadastral building sub-domain.

Though each source system has its own data schema, there is no common data schema available for this data integration process. The next step is to define rules and exceptions to build a common data schema on the class and attribute levels.

Rules to specify a common data schema on the class levels. This type of rule decides which source system is the master for a specified class. For example, the "Organization" class from the Business Entity Register and the "Building" class from the cadastral system are the master classes with national unique identifiers. However there are also exceptions because of some special business rules in practice. For example: Buildings less than 15 square meters are not required to be registered in the cadastral register, neither do the embassy buildings in foreign countries. A supplementary unique identifier for the "Building" class is needed to handle the exception buildings without national unique identifier from the cadastral system.

Rules to specify a common data schema on the attribute levels. The rules decide which source system is the master for some specified attributes. For example: the Building Accessibility Register is the master for the accessibility attributes of a building though the "Building" class in the cadastral system is defined as the master class.

The last step in this sub-section defines rules to specify attributes that can be used to connect heterogeneous data sources (integration keys). The integration keys are normally the unique identifiers of the master classes. For example, the organization number for a real estate owner is an integration key to connect the cadastral system to Business Entity Register. There are exceptions in cases such as a supplementary unique identifier is needed to cover buildings less than 15 square meters. Both the primary and supplementary unique identifiers are used in the integration to return a complete building list for the SoE report service.

3.2 Rules for data quality assessment

The result data is an integrated result of multiple source data. Both the source and result data should be screened for potential syntactic and semantic errors using data quality rules generated from existing domain models or expert knowledge. Examples of different types of data quality rules include:

- *Mandatory:* The property owner is mandatory for a property ownership record. The rule is broken when the property owner is missing.
- *Data type:* The area field of a building should be numeric. The rule is broken when the area field set to text "N/A".
- *Data length:* A municipality number should be four digits. The rule is broken when a municipality number is made of three digits.
- *Uniqueness:* The cadastral building number should be unique. It breaks the rule when one cadastral building number is registered on more than one building in the Statsbygg's property management system.
- *Cardinality:* A cadastral parcel is located in a municipality. The rule is broken if the municipality field is missing.
- *Data domain and range:* The valid values of cadastral parcel ownership types in this case should be either owned or leased. Including other values than those two breaks the rule.

3.3 Measurement metrics for data integration

In addition to the above rules, Table 1 shows three measurement metrics that are dedicated to identify quality problems in the data integration and measure the quality of the integration result.

The integration keys are attributes used to connect heterogeneous data sources, and they are currently registered manually in the referring systems (systems that refer to the key attributes). There is no automatic updating of the values of integration keys after registration. For example the organization number is registered as an identifier for property owners when an ownership record is created in the cadastral system and it is then kept to be static in the cadastral system and does not follow the changes or

deletions in Business Entity Register. Some of the integration keys are not mandatory fields in the referring systems. Here are two examples of the Key Value Quality metrics: 1) the percentage of outdated organization numbers for the property owners in the cadastral system; 2) the percentage of missing or outdated cadastral building numbers in the Statsbygg's property management system.

Table 1. Measurement metrics type on data integration

Metrics type	Description
Key Value Quality	The number or percentage of missing or outdated key values registered in source data
Integration Quality	The number or percentage of incorrect integrations in the result data.
Non-matched rows	The number of non-matched rows in an integration identified using an outer join or similar techniques.

The Integration Quality metric measures the percentage of correct integrations in the integration result. Though the key values may exist in the source system for master data, it could also refer to a wrong data item. The cadastral ownership data contains both the property owners' organization number and name. The organization number is used as an integration key to integrate the cadastral ownership data with Business Entity Register. We identify afterwards the deviation between organization names from two data sources to measure the correctness of the integration result.

The non-matched rows metric returns the rows from one system that is not able to be integrated with another system. This metric is especially useful when a supplementary source data has partly more updated information on some specified data items than the primary source data.

3.4 Strategies for improving data quality

Examples of measurement metrics and corresponding quality improvement strategies are presented in Table 2. The result of data integration [4] for the SoE service is provisioned as RDF/Linked Data through a Linked Data generation process supported by DataGraft[10] [5][6][7] and is available via the proDataMarket marketplace[11]. SPARQL queries have been used as the underlying mechanism to assess the data quality. The query results help the responsible staff assess and improve the data quality in the source systems by following the suggested strategy for quality improvement. The updated source data with better data quality are then reloaded to the integration process to produce an updated result with improved quality.

[10] https://datagraft.io/

[11] https://prodatamarket.eu/

Table 2. Examples of measurement metrics and quality improvement actions

SPARQL query to identify:	Type of measurement metrics	Possible reasons of mismatch	Strategy to improve data quality
The owner name difference between cadastral system and business entity register[12]	Integration Quality	Delayed or missing updates of owner names in the cadastral system.	Update the owner names in the cadastral system.
The state-owned properties that are missing in the previous SoE report[13]	Non-matched rows	The properties were acquired after the previous report was made.	No specific actions needed though it reflects partially the quality of the previous SoE report.
		The properties were forgotten to be registered in the previous SoE report.	
The state-owned properties from the previous SoE report that are missing in the resulting SoE report[14]	Non-matched rows	The properties were sold to a non-central government organization after the previous report was made.	No specific actions needed.
		The properties are abroad.	
		There has been organization change with the owner and the owner's organization number is no longer valid in the business entity register.	Update the owner's organization number and name in the cadastral system.
		The ownership change between organizations in the public sector is not always officially registered in the cadastral system.	Inform the current owner organization to update the ownership in the cadastral system.
		The owner's organization is not officially registered as central government organization in the business entity register.	Update the organization in the business entity register if it is applicably or add it to the manual exception list of the central government organization data.

[12] https://datagraft.io/prodatamarket_publisher/queries/soe-query1-the-owner-name-difference

[13] https://datagraft.io/prodatamarket_publisher/queries/soe-query2-missing-soe-records

[14] https://datagraft.io/prodatamarket_publisher/queries/soe-query3-missing-result-soe-records

4 Summary and outlook

This paper introduced the State of Estate report service together with its value proposition and the rule-based approach to address data quality issues. The report service is a result of integrating cross-domain data from multiple sources as the cadastral system, Business Entity Register, Building Accessibility Register and Statsbygg's property management system. A set of rules are developed to meet the data quality requirements on SoE report service, including rules to specify a common data schema, rules for data quality assessment and measurement metrics for data integration. Strategies for improving data quality are also presented. The rule-based approach is quick to implement and easily understandable both by domain experts and data engineers.

For the further work, the identified rules shall be transformed to executable rules if possible such that they can be applied directly in semantic reasoning to automate the quality assessment process. The suggested quality improvement strategies can also be half or fully automated to increase effectivity.

Acknowledgements. The work in this paper is partly supported by the EC funded project proDataMarket (Grant number: 644497).

References

1. Shi, L., Pettersen, B. E., Østhassel, I., Nikolov, N., Khorramhonarnama, A., Berre, A. J., & Roman, D. (2015, August). Norwegian State of Estate: A Reporting Service for the State-Owned Properties in Norway. In International Symposium on Rules and Rule Markup Languages for the Semantic Web (pp. 456-464). Springer International Publishing.
2. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211-218.
3. Chiang, F., & Miller, R. J. (2008). Discovering data quality rules. *Proceedings of the VLDB Endowment*, *1*(1), 1166-1177.
4. Halevy, A., Rajaraman, A., & Ordille, J. (2006, September). Data integration: the teenage years. In *Proceedings of the 32nd international conference on Very large data bases* (pp. 9-16). VLDB Endowment.
5. Roman, D., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A. J., Ye, X., Dimitrov, M., Simov, A., Zarev, M., Moynihan, R., Roberts, B., Berlocher, I., Kim, S., Lee, T., Smith, A., & Heath, T. DataGraft: One-Stop-Shop for Open Data Management. To appear in the *Semantic Web Journal (SWJ) – Interoperability, Usability, Applicability* (published and printed by IOS Press, ISSN: 1570-0844), 2017, DOI: 10.3233/SW-170263.
6. Roman, D., Dimitrov, M., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A. J., Ye, X., Simov, A. & Petkov, Y. DataGraft: Simplifying Open Data Publishing. *ESWC (Satellite Events) 2016: 101-106*.
7. Roman, D., Dimitrov, M., Nikolov, N., Putlier, A., Elvesæter, B., Simov, A., Petkov, Y. DataGraft: A Platform for Open Data Publishing. In the *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. (LIME/SemDev@ESWC 2016)*.